Multi-Modal Representations for Improved Bilingual Lexicon Learning

Ivan Vulić
Language Technology Lab, DTAL
University of Cambridge
iv250@cam.ac.uk

Douwe Kiela
Computer Laboratory
University of Cambridge
dk427@cam.ac.uk

Stephen Clark
Computer Laboratory
University of Cambridge
sc609@cam.ac.uk

Marie-Francine Moens
Department of Computer Science
KU Leuven
sien.moens@cs.kuleuven.be

Abstract

Recent work has revealed the potential of using visual representations for bilingual lexicon learning (BLL). Such image-based BLL methods, however, still fall short of linguistic approaches. In this paper, we propose a simple yet effective multi-modal approach that learns bilingual semantic representations that fuse linguistic and visual input. These new bilingual multi-modal embeddings display significant performance gains in the BLL task for three language pairs on two benchmarking test sets, outperforming linguistic-only BLL models using three different types of state-of-the-art bilingual word embeddings, as well as visual-only BLL models.

1 Introduction

Bilingual lexicon learning (BLL) is the task of finding words that share a common meaning across different languages. It plays an important role in a variety of fundamental tasks in IR and NLP, e.g. cross-lingual information retrieval and statistical machine translation. The majority of current BLL models aim to learn lexicons from comparable data. These approaches work by (1) mapping language pairs to a *shared cross-lingual vector space (SCLVS)* such that words are close when they have similar meanings; and (2) extracting close lexical items from the induced SCLVS. Bilingual word embedding (BWE) induced models currently hold the state-of-the-art on BLL (Hermann and Blunsom, 2014; Gouws et al., 2015; Vulić and Moens, 2016).

Although methods for learning SCLVSs are predominantly text-based, this space need not be linguistic in nature: Bergsma and van Durme (2011) and Kiela et al. (2015) used labeled images from the Web to learn bilingual lexicons based on *visual* features, with features derived from deep convolutional neural networks (CNNs) leading to the best results (Kiela et al., 2015). However, vision-based BLL does not yet perform at the same level as state-of-the-art linguistic models. Here, we unify the strengths of both approaches into one single multi-modal vision-language SCLVS.

It has been found in multi-modal semantics that linguistic and visual representations are often complementary in terms of the information they encode (Deselaers and Ferrari, 2011; Bruni et al., 2014; Silberer and Lapata, 2014). This is the first work to test the effectiveness of the multi-modal approach in a BLL setting. Our contributions are: We introduce bilingual multi-modal semantic spaces that merge linguistic and visual components to obtain semantically-enriched bilingual multi-modal word representations. These representations display significant improvements for three language pairs on two benchmarking BLL test sets in comparison to three different bilingual linguistic representations (Mikolov et al., 2013; Gouws et al., 2015; Vulić and Moens, 2016), as well as over the uni-modal visual representations from Kiela et al. (2015).

We also propose a weighting technique based on image dispersion (Kiela et al., 2014) that governs the influence of visual information in fused representations, and show that this technique leads to robust multi-modal models which do not require fine tuning of the fusion parameter.

2 Methodology

2.1 Linguistic Representations

We use three representative linguistic BWE models. Given a source and target vocabulary V^S and V^T, BWE models learn a representation of each word $w \in V^S \cup V^T$ as a real-valued vec-

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 188–194,
Berlin, Germany, August 7-12, 2016.

tor: $\mathbf{w}_{ling} = [f_1^{ling}, \ldots, f_{d_l}^{ling}]$, where $f_k^{ling} \in \mathbb{R}$ is the value of the k-th cross-lingual feature for w. Similarity between $w, v \in V^S \cup V^T$ is computed through a similarity function (SF), $sim_{ling}(w, v) = SF(\mathbf{w}_{ling}, \mathbf{v}_{ling})$, e.g., cosine.

Type 1: M-EMB This type of BWE induction model assumes the following setup for learning the SCLVS (Mikolov et al., 2013; Faruqui and Dyer, 2014; Dinu et al., 2015; Lazaridou et al., 2015a): First, two monolingual spaces, $\mathbb{R}^{d_S}$ and $\mathbb{R}^{d_T}$, are induced separately in each language using a standard monolingual embedding model. The bilingual signal is provided in the form of word translation pairs (x_i, y_i), where $x_i \in V^S$, $y_i \in V^T$, and $\mathbf{x_i} \in \mathbb{R}^{d_S}$, $\mathbf{y_i} \in \mathbb{R}^{d_T}$. Training is cast as a multivariate regression problem: it implies learning a function that maps the source language vectors to their corresponding target language vectors. A standard approach (Mikolov et al., 2013; Dinu et al., 2015) is to assume a linear map $\mathbf{W} \in \mathbb{R}^{d_S \times d_T}$, which is learned through an L_2-regularized least-squares error objective. Any previously unseen source language word vector $\mathbf{x_u}$ may be mapped into the target embedding space $\mathbb{R}^{d_T}$ as $\mathbf{W}\mathbf{x_u}$. After mapping all vectors $\mathbf{x}$, $x \in V^S$, the target space $\mathbb{R}^{d_T}$ serves as a SCLVS.

Type 2: G-EMB Another collection of BWE induction models optimizes two monolingual objectives *jointly*, with the cross-lingual objective acting as a cross-lingual regularizer during training (Gouws et al., 2015; Soyer et al., 2015). In a simplified formulation (Luong et al., 2015), the objective is: $\gamma(\mathit{Mono}_S + \mathit{Mono}_T) + \delta \mathit{Bi}$. The monolingual objectives Mono_S and Mono_T ensure that similar words in each language are assigned similar embeddings and aim to capture the semantic structure of each language, whereas the cross-lingual objective Bi ensures that similar words across languages are assigned similar embeddings, and ties the two monolingual spaces together into a SCLVS. Parameters γ and δ govern the influence of the monolingual and bilingual components.[1] The bilingual signal used as the cross-lingual regularizer during the joint training is obtained from sentence-aligned parallel data. We opt for the BilBOWA model from Gouws et al. (2015) as the representative model to be included in the comparisons, due to its solid performance and robustness in the BLL task (Luong et al., 2015), its reduced complexity reflected in fast computations on massive datasets and its public availability.[2]

Type 3: V-EMB The third set of models requires a different bilingual signal to induce a SCLVS: *document alignments*. Vulić and Moens (2016) created a collection of pseudo-bilingual documents by merging every pair of aligned documents in the data, in a way that preserves important local information – which words appeared next to which other words (in the same language), and which words appeared in the same region of the document (in different languages). This collection was then used to train word embeddings with monolingual skip-gram with negative sampling using `word2vec`. With pseudo-bilingual documents, the "context" of a word is redefined as a mixture of neighboring words (in the original language) and words that appeared in the same region of the document (in the foreign language). Bilingual contexts for each word in each pseudo-bilingual document steer the final model towards constructing a SCLVS.

2.2 Visual Representations

Only a few studies have tried to make use of the intuition that words in different languages denoting the same concepts are similarly grounded in the perceptual system (bicycles resemble each other irrespective of whether we call them *bicyle*, *vélo*, *fiets* or *Fahrrad*, see Fig. 1) (Bergsma and van Durme, 2011; Kiela et al., 2015). Although the idea is promising, such visual methods are still limited in comparison with linguistic ones, especially for more abstract concepts (Kiela et al., 2015). Recent findings in multi-modal semantics suggest that visual representations encode pieces of semantic information complementary to linguistic information derived from text (Deselaers and Ferrari, 2011; Silberer and Lapata, 2014).

We compute visual representations in a similar fashion to Kiela et al. (2015): For each word we retrieve n images from Google image search (see Fig. 1), and for each image we extract the pre-softmax layer of an AlexNet (Krizhevsky et al., 2012) that has been pre-trained on the ImageNet

[1] Setting $\gamma = 0$ reduces the model to the bilingual models trained solely on parallel data (Hermann and Blunsom, 2014; Chandar et al., 2014). $\gamma = 1$ results in the models from Gouws et al. (2015) and Soyer et al. (2015). Although they use the same data sources, all G-EMB models differ in the choice of monolingual and cross-lingual objectives.

[2] https://github.com/gouwsmeister/bilbowa

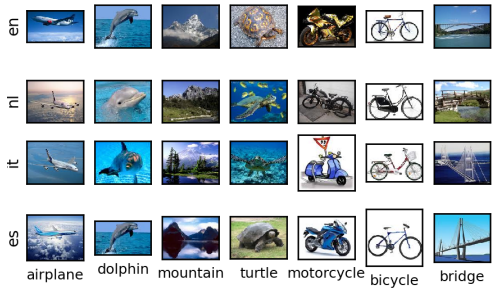

Figure 1: Example images for several languages.

classification task (Deng et al., 2009; Russakovsky et al., 2015) using Caffe (Jia et al., 2014).

Each image is thus represented as a 4096-dimensional feature vector extracted from a convolutional neural network (CNN). We use two methods for computing visual similarity: (1) CNN-MAX produces a single visual vector by taking the pointwise maximum across the n image vector representations from the image set. The representation of each word $w \in V^S \cup V^T$ in a visual SCLVS is now a real-valued vector $\mathbf{w}_{vis} = [f_1^{vis}, \ldots, f_{d_v}^{vis}]$, where $f_k^{vis} \in \mathbb{R}$ denotes the score for the k-th visual cross-lingual feature for w within a d_v-dimensional visual SCLVS ($d_v = 4096$). As before, similarity between two words $w, v \in V^S \cup V^T$ is computed by applying a similarity function on their representations in the visual SCLVS: $sim_{vis}(w, v) = SF(\mathbf{w}_{vis}, \mathbf{v}_{vis})$, e.g. cosine. (2) CNN-AVGMAX: An alternative strategy, introduced by Bergsma and van Durme (2011), is to consider the similarities between individual images from the two sets and take the average of the maximum similarity scores as the final similarity $sim_{vis}(w, v)$.

2.3 Multi-Modal Representations

We experiment with two ways of fusing information stemming from the linguistic and visual modalities. Following recent work in multi-modal semantics (Bruni et al., 2014; Kiela and Bottou, 2014), we construct representations by concatenating the centered and L_2-normalized linguistic and visual feature vectors:

$$\mathbf{w}_{mm} = \alpha \times \mathbf{w}_{ling} \;||\; (1-\alpha) \times \mathbf{w}_{vis} \qquad (1)$$

where $||$ denotes concatenation and α is a parameter governing the contributions of each uni-modal representation. The final similarity may again be computed by applying an SF on the multi-modal representations. We call this method **Early-Fusion**. Note that it is possible only with CNN-MAX. The alternative is not to build a full multi-modal (MM) representation, but instead to combine the individual similarity scores from each uni-modal SCLVS. The similarity $sim(w, v)$ between two words w and v is:

$$\alpha \times sim_{ling}(w, v) + (1-\alpha) \times sim_{vis}(w, v) = \\ = \alpha \times SF(\mathbf{w}_{ling}, \mathbf{v}_{ling}) + (1-\alpha) \times SF(\mathbf{w}_{vis}, \mathbf{v}_{vis})$$

where α again controls for the importance of the uni-modal scores in the final combined scores. We call this method **Late-Fusion**[3].

3 Experimental Setup

Task: Bilingual Lexicon Learning Given a source language word w_s, the task is to find a target language word w_t closest to w_s in the SCLVS, and the resulting pair (w_s, w_t) is a bilingual lexicon entry. Performance is measured using the BLL standard *Top 1* accuracy (Acc_1) metric (Gaussier et al., 2004; Gouws et al., 2015).

Test Sets We work with three language pairs: English-Spanish/Dutch/Italian (EN-ES/NL/IT), and two benchmarking BLL test sets:
(1) BERGSMA500: consisting of a set of 500 ground truth noun pairs for the three language pairs, it is considered a benchmarking test set in prior work on BLL using vision (Bergsma and van Durme, 2011)[4]. Translation direction in our tests is $EN \rightarrow ES/IT/NL$.
(2) VULIC1000: constructed to measure the general performance of linguistic BLL models from comparable Wikipedia data (Vulić and Moens, 2013), this is considered a benchmarking test set for (linguistic) BLL models from comparable data (Vulić and Moens, 2016)[5]. It comprises 1,000 nouns in ES, IT, and NL, along with their one-to-one ground-truth word translations in EN compiled semi-automatically. Translation direction is $ES/IT/NL \rightarrow EN$.

Training Data and Setup We used standard training data and suggested settings to learn M/G/V-EMB model representations. M-EMB and G-EMB were trained on the full cleaned and tokenized Wikipedias from the Polyglot website (Al-Rfou et al., 2013). V-EMB was trained on the full tokenized document-aligned Wikipedias from

[3]Under the assumption of having the centered and L_2-normalized feature vectors, and *cos* as SF, Early-Fusion may be transformed into Late-Fusion with adapted weighting:
$\alpha^2 \times cos(\mathbf{w}_{ling}, \mathbf{v}_{ling}) + (1-\alpha)^2 \times cos(\mathbf{w}_{vis}, \mathbf{v}_{vis})$
[4]http://www.clsp.jhu.edu/~sbergsma/LexImg/
[5]http://www.cl.cam.ac.uk/~dk427/bli.html

Pair:	B: EN→ES|*V: ES→EN*						B: EN→IT|*V: IT→EN*						B: EN→NL|*V: NL→EN*					
Models	M-EMB		G-EMB		V-EMB		M-EMB		G-EMB		V-EMB		M-EMB		G-EMB		V-EMB	
Linguistic																		
$d = 300$	0.71	*0.77*	0.60	*0.73*	0.68	*0.82*	0.77	*0.76*	0.63	*0.71*	0.75	*0.79*	0.77	*0.76*	0.59	*0.75*	0.74	*0.79*
Visual																		
CNN-Max	0.51	*0.35*	0.51	*0.35*	0.51	*0.35*	0.54	*0.22*	0.54	*0.22*	0.54	*0.22*	0.56	*0.33*	0.56	*0.33*	0.56	*0.33*
CNN-AvgMax	0.55	*0.38*	0.54	*0.38*	0.54	*0.38*	0.56	*0.25*	0.56	*0.25*	0.56	*0.25*	0.60	*0.34*	0.60	*0.34*	0.60	*0.34*
Multi-modal with global α																		
Max-E-0.5	0.76	*0.79*	0.66	***0.79***	0.71	*0.83*	0.83	*0.75*	0.72	*0.70*	0.80	*0.80*	0.85	*0.80*	0.69	*0.78*	0.80	*0.81*
Max-E-0.7	0.75	*0.80*	0.62	*0.76*	0.70	***0.85***	0.81	*0.77*	0.66	*0.73*	0.78	*0.82*	0.84	*0.80*	0.61	*0.79*	0.80	*0.82*
Max-L-0.7	0.76	*0.80*	0.64	*0.78*	0.71	***0.85***	0.82	*0.77*	0.69	*0.73*	0.80	*0.82*	0.85	*0.82*	0.64	*0.79*	0.81	***0.83***
Avg-L-0.5	**0.77**	*0.78*	**0.68**	***0.79***	**0.73**	*0.83*	**0.84**	*0.77*	**0.75**	*0.70*	**0.81**	*0.79*	**0.86**	*0.80*	**0.76**	*0.78*	**0.83**	*0.81*
Avg-L-0.7	**0.77**	***0.81***	0.66	***0.79***	0.72	***0.85***	0.83	***0.78***	0.72	***0.75***	0.80	***0.83***	**0.86**	***0.83***	0.70	***0.81***	0.81	***0.83***
Multi-modal with image dispersion (ID) weighting																		
Max-E-ID	0.76	*0.80*	0.66	*0.78*	0.71	*0.84*	0.81	*0.77*	0.69	*0.73*	0.80	*0.81*	0.84	*0.80*	0.64	*0.79*	0.81	*0.82*
Max-L-ID	**0.77**	*0.80*	0.66	*0.78*	0.72	***0.85***	0.82	*0.77*	0.70	*0.73*	0.80	*0.81*	0.84	*0.82*	0.65	*0.79*	0.81	*0.82*
Avg-L-ID	**0.77**	***0.81***	0.67	***0.79***	**0.73**	*0.84*	0.83	***0.78***	0.74	*0.73*	0.80	***0.83***	0.85	*0.82*	0.72	*0.80*	0.82	*0.82*

Table 1: Summary of the Acc_1 scores on BERGSMA500 (regular font) and VULIC1000 (*italic*) across all BLL runs. M/G/V-EMB denotes the BWE linguistic model. Other settings are in the form Y-Z-0.W: (1) Y denotes the visual metric, (2) Z denotes the fusion model: E is for Early-Fusion, L is for Late-Fusion, and (3) 0.W denotes the α value. Highest scores per column are in bold.

LinguaTools[6]. The 100K most frequent words were retained for all models.

We followed related work (Mikolov et al., 2013; Lazaridou et al., 2015a) for learning the mapping $\mathbf{W}$ in M-EMB: starting from the BNC word frequency list (Kilgarriff, 1997), the 6,318 most frequent EN words were translated to the three other languages using Google Translate. The lists were subsequently cleaned, removing all pairs that contain IT/ES/NL words occurring in the test sets and least frequent pairs, to build the final 3×5K training pairs. We trained two monolingual SGNS models, using SGD with a global learning rate of 0.025. For G-EMB, as in the original work (Gouws et al., 2015), the bilingual signal for the cross-lingual regularization was provided in the first 500K sentences from Europarl.v7 (Tiedemann, 2012). We used SGD with a global learning rate 0.15. For V-EMB, monolingual SGNS was trained on pseudo-bilingual documents using SGD with a global learning rate 0.025. All BWEs were trained with $d = 300$.[7] Other parameters are: 15 epochs, 15 negatives, subsampling rate $1e-4$. We report results with two α standard values: 0.5 and 0.7 (more weight assigned to the linguistic part).

4 Results and Discussion

Table 1 summarizes Acc_1 scores, focusing on interesting comparisons across different dimensions[8]. There is a marked difference in performance on BERGSMA500 and VULIC1000: visual-only BLL models on VULIC1000 perform two times worse than linguistic-only BLL models. This is easily explained by the increased abstractness of test words in VULIC1000 in comparison to BERGSMA500[9], which highlights the need for a multi-modal approach.

Multi-Modal vs. Uni-Modal The multi-modal models outperform both linguistic and visual models across all setups and combinations on BERGSMA500. On VULIC1000 multi-modal models again outperform their uni-modal components in both modalities. In the latter case, improvements are dependent on the amount of visual information included in the model, as governed by α. Since the dataset also contains highly abstract words, the inclusion of visual information may be detrimental to performance. These models outperform the uni-modal models across a wide variety of settings: they outperform the three linguistic-only BLL models that held best reported Acc_1 scores on the evaluation set (Vulić and Moens, 2016). The largest improvements are statistically significant according to McNemar's test, $p < 0.01$. We find improvements on both test sets for all three BWE types.

The relative ranking of the visual metrics intro-

[6] http://linguatools.org/tools/corpora/

[7] Similar trends were observed with all models and $d = 64, 500$. We also vary the window size from 4 to 16 in steps of 4, and always report the best scoring linguistic embeddings.

[8] Similar rankings of different models are also visible with more lenient Acc_{10} scores, not reported for brevity.

[9] The average image dispersion value (Kiela et al., 2014), which indicates abstractness, on VULIC1000 is 0.711 compared to 0.642 on BERGSMA500.

duced in Kiela et al. (2015) extends to the MM setting: Late-Fusion with CNN-AVGMAX is the most effective MM BLL model on average, but all other tested MM configurations also yield notable improvements.

Concreteness To measure concreteness, we use an unsupervised data-driven method, shown to closely mirror how concrete a concept is: *image dispersion (ID)* (Kiela et al., 2014). ID is defined as the average pairwise cosine distance between all the image representations/vectors $\{\mathbf{i}_1 \dots \mathbf{i}_n\}$ in the set of images for a given word w:

$$id(w) = \frac{2}{n(n-1)} \sum_{j<k\leq n} 1 - \frac{\mathbf{i}_j \cdot \mathbf{i}_k}{|\mathbf{i}_j||\mathbf{i}_k|} \quad (2)$$

Intuitively, more concrete words display more coherent visual representations and consequently lower ID scores (see Footnote 9 again). The lowest improvements on VULIC1000 are reported for the IT-EN language pair, which is incidentally the most abstract test set.

There is some evidence that abstract concepts are also perceptually grounded (Lakoff and Johnson, 1999), albeit in a more complex way, since abstract concepts will relate more varied situations (Barsalou and Wiemer-Hastings, 2005). Consequently, uni-modal visual representations are not powerful enough to capture all the semantic intricacies of such abstract concepts, and the linguistic components are more beneficial in such cases. This explains an improved performance with $\alpha = 0.7$, but also calls for a more intelligent decision mechanism on how much perceptual information to include in the multi-modal models. The decision should be closely related to the degree of a concept's concreteness, e.g., eq. (2).

Image Dispersion Weighting The intuition that the inclusion of visual information may lead to negative effects in MM modeling has been exploited by Kiela et al. (2014) in their work on image-dispersion filtering: Although the filtering method displays some clear benefits, its shortcoming lies in the fact that it performs a binary decision which can potentially discard valuable perceptual information for less concrete concepts. Here, we introduce a weighting scheme where the perceptual information is weighted according to its ID value. Early-Fusion is now computed as:

$$\mathbf{w}_{mm} = \alpha(id) \times \mathbf{w}_{ling} \;||\; (1 - \alpha(id)) \times \mathbf{w}_{vis}$$

Late-Fusion model becomes:

$$\alpha(id) \times SF(\mathbf{w}_{ling}, \mathbf{v}_{ling}) \; + (1 - \alpha(id)) \times SF(\mathbf{w}_{vis}, \mathbf{v}_{vis})$$

$\alpha(id)$ denotes a weight that is proportional to the ID score of the source language word w: we opt for a simple approach and specify $\alpha(id) = id(w)$. Instead of having one global parameter α, the ID weighting adjusts the amount of information locally according to each concept's concreteness.

The results are summarised in Table 1. All multi-modal models with ID-based weighting are outperforming their uni-modal components. The ID-weighted BLL models reach (near-)optimal BLL results across a variety of language-vision combinations without any fine-tuning.

5 Conclusion

We have presented a novel approach to bilingual lexicon learning (BLL) that combines linguistic and visual representations into new bilingual multi-modal (MM) models. Two simple yet effective ways to fuse the linguistic and visual information for BLL have been described. Such MM models outperform their linguistic and visual uni-modal component models on two standard benchmarking BLL test sets for three language pairs. Comparisons with three different state-of-the-art bilingual word embedding induction models demonstrate that the gains of MM modeling are generally applicable.

As future work, we plan to analyse the ability of multi-view representation learning algorithms to yield fused multi-modal representations in bilingual settings (Lazaridou et al., 2015b; Rastogi et al., 2015; Wang et al., 2015), as well as to apply multi-modal bilingual spaces in other tasks such as zero-short learning (Frome et al., 2013) or cross-lingual MM information search and retrieval following paradigms from monolingual settings (Pereira et al., 2014; Vulić and Moens, 2015).

The inclusion of perceptual data, as this paper reveals, seems especially promising in bilingual settings (Rajendran et al., 2016; Elliott et al., 2016), since the perceptual information demonstrates the ability to transcend linguistic borders.

Acknowledgments

This work is supported by ERC Consolidator Grant LEXICAL (648909) and KU Leuven Grant PDMK/14/117. SC is supported by ERC Starting Grant DisCoTex (306920). We thank the anonymous reviewers for their helpful comments.

References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. In *CoNLL*, pages 183–192.

Lawrence W. Barsalou and Katja Wiemer-Hastings. 2005. Situating abstract concepts. In D. Pecher and R. Zwaan, editors, *Grounding cognition: The role of perception and action in memory, language, and thought*, pages 129–163.

Shane Bergsma and Benjamin van Durme. 2011. Learning bilingual lexicons using the visual similarity of labeled web images. In *IJCAI*, pages 1764–1769.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artiifical Intelligence Research*, 49:1–47.

Sarath A.P. Chandar, Stanislas Lauly, Hugo Larochelle, Mitesh M. Khapra, Balaraman Ravindran, Vikas C. Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *NIPS*, pages 1853–1861.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255.

Thomas Deselaers and Vittorio Ferrari. 2011. Visual and semantic similarity in ImageNet. In *CVPR*, pages 1777–1784.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *ICLR Workshop Papers*.

D. Elliott, S. Frank, K. Sima'an, and L. Specia. 2016. Multi30K: Multilingual English-German Image Descriptions. *CoRR*, abs/1605.00459.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *EACL*, pages 462–471.

Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129.

Éric Gaussier, Jean-Michel Renders, Irina Matveeva, Cyril Goutte, and Hervé Déjean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *ACL*, pages 526–533.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. BilBOWA: Fast bilingual distributed representations without word alignments. In *ICML*, pages 748–756.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *ACL*, pages 58–68.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross B. Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia*, pages 675–678.

Douwe Kiela and Léon Bottou. 2014. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *EMNLP*, pages 36–45.

Douwe Kiela, Felix Hill, Anna Korhonen, and Stephen Clark. 2014. Improving multi-modal representations using image dispersion: Why less is sometimes more. In *ACL*, pages 835–841.

Douwe Kiela, Ivan Vulić, and Stephen Clark. 2015. Visual bilingual lexicon induction with transferred ConvNet features. In *EMNLP*, pages 148–158.

Adam Kilgarriff. 1997. Putting frequencies in the dictionary. *International Journal of Lexicography*, 10(2):135–155.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114.

George Lakoff and Mark Johnson. 1999. *Philosophy in the flesh: The embodied mind and its challenge to Western thought*.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015a. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *ACL*, pages 270–280.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015b. Combining language and vision with a multimodal skip-gram model. In *NAACL-HLT*, pages 153–163.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert R. G. Lanckriet, Roger Levy, and Nuno Vasconcelos. 2014. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):521–535.

Janarathanan Rajendran, Mitesh M. Kapra, Sarath Chandar, and Balaraman Ravindran. 2016. Bridge correlational neural networks for multilingual multimodal representation learning. In *NAACL*.

Pushpendre Rastogi, Benjamin Van Durme, and Raman Arora. 2015. Multiview LSA: Representation learning via generalized CCA. In *NAACL*, pages 556–566.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *ACL*, pages 721–732.

Hubert Soyer, Pontus Stenetorp, and Akiko Aizawa. 2015. Leveraging monolingual data for crosslingual compositional word representations. In *ICLR*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *LREC*, pages 2214–2218.

Ivan Vulić and Marie-Francine Moens. 2013. A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else). In *EMNLP*, pages 1613–1624.

Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *SIGIR*, pages 363–372.

Ivan Vulić and Marie-Francine Moens. 2016. Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research*, 55:953–994.

Weiran Wang, Raman Arora, Karen Livescu, and Jeff A. Bilmes. 2015. On deep multi-view representation learning. In *ICML*, pages 1083–1092.